WHAT IS CLAIMED IS:

1. In a data processing system, a method for creating a database from information found on a plurality of web pages using a first classifier and a second classifier, said method comprising:

a) defining first regularities and second regularities, said first regularities being patterns which are expected to be found in information in said web pages, and said second regularities being patterns which are not expected to be found in all said web pages;

b) initially providing descriptions of said first regularities to a working database; thereafter

c) training a first classifier in said working database using said first regularities;

d) identifying a candidate subset of the web pages expected to have said second regularities; thereafter

e) tentatively identifying and tagging, in said candidate subset of the web pages, elements having said first regularities, by using said first classifier to obtain first tentative labels;

f) training a second classifier using said first tentative labels;

g) tentatively identifying elements having specific combinations of said first regularities and said second regularities using said first classifier and said second classifier to obtain second tentative labels for said elements of said candidate subset; thereafter

h) outputting, said second tentative labels as permanent labels associated with said elements of said candidate subset of web pages.

2. The method of claim 1 further including:

h) deciding whether to retrain said second classifier with said second tentative labels.

3. The method according to claim 2, further including:

f) training the second classifier using said second tentative labels.

4. The method according to claim 2, further including

g) collecting said permanent labels associated with said elements of said candidate subset of web pages;

h) retraining said first classifier in response to said permanent labels.

5. The method according to claim 1 wherein said second classifier treats selected first regularities differently than said first classifier treats said first regularities such that said second regularities contradict said first regularities.

6. The method according to claim 5 wherein said outputting step further includes ignoring training results of said first classifier.

7. The method according to claim 5 wherein said outputting step further includes combining training results of said first classifier and said second classifier.

8. In a data processing system, a method for learning and combining global regularities and local regularities for information extraction and classification, said method comprising the steps of

a) initially providing descriptions of said global regularities to a working database, said global regularities being patterns which may be found over an entire dataset; thereafter

b) identifying a candidate subset of the dataset in which local regularities may be found; thereafter

c) tentatively identifying elements having said global regularities in said candidate subset to obtain first tentative labels, said first tentative labels being useful for tagging information having identifiable similarities; thereafter

d) attaching said first tentative labels onto said identified elements of said candidate subset; thereafter

e) employing said attached first tentative labels via one of a class of inductive operations to formulate first local regularities; thereafter

f) tentatively identifying elements having specific combinations of said global regularities and said local regularities to obtain attached second tentative labels; thereafter

g) testing if estimated error rate is within a preselected tolerance or if a steady state in said attached second tentative labels is evident; and if true, then rating

confidence of said attached second tentative labels and converting selected ones of said attached second tentative labels to confidence labels upon achieving a preselected confidence level and then outputting data with said confidence labels; otherwise

h) employing said second tentative labels via said operation on said candidate subset to formulate second local regularities, and

i) repeating from step f) until said confidence labels have been developed.

9. The method according to claim 8 wherein said initial global regularity providing step comprises manually inputting descriptions of said global regularities.

10. The method according to claim 8 wherein said initial global regularity providing step comprises obtaining said global regularities from a further one of said class of said inductive operations applied to a subset of said dataset, said subset of said dataset having been manually labeled.

11. The method according to claim 10 further including developing refined global regularities comprising the steps of:

l) collecting confidence labels from at least one of said candidate subsets to obtain global confidence labels;

m) employing said global confidence labels on candidate subsets along with said manually labeled dataset via one of said class of inductive operations to formulate said refined global regularities;

n) providing descriptions of said refined global regularities to said working database; thereafter

o) identifying a next candidate subset of the dataset in which local regularities may be found; thereafter

p) tentatively identifying elements having said refined global regularities in said candidate subset to obtain next tentative labels; thereafter

q) attaching said next tentative labels onto said identified elements of said next candidate subset; thereafter

r) employing said attached next tentative labels via one of the class of inductive operations to formulate next local regularities; thereafter

s) tentatively identifying elements having specific combinations of said refined global regularities and said next local regularities to obtain attached next second tentative labels; thereafter

t) testing if estimated error rate is within a preselected tolerance or if a steady state in said next second tentative labels is evident; and if true, then rating confidence of said attached next second tentative labels and converting selected ones of said attached next second tentative labels to confidence labels upon achieving a preselected confidence level and then outputting data with said confidence labels;

u) otherwise employing said next second tentative labels via said operation on said candidate subset to formulate next second local regularities, and

v) repeating from step s).

12. The method according to claim 11 further including the steps of:

applying said data with confidence labels to further subsets of said dataset to investigate further subsets for local regularities.

13. The method according to claim 8 further including developing refined global regularities comprising the steps of:

l) collecting confidence labels from at least one of said candidate subsets to obtain global confidence labels;

m) employing said global confidence labels on candidate subsets via one of said class of inductive operations to formulate said refined global regularities;

n) providing descriptions of said refined global regularities to said working database; thereafter

o) identifying a next candidate subset of the dataset in which local regularities may be found; thereafter

p) tentatively identifying elements having said refined global regularities in said candidate subset to obtain next tentative labels; thereafter

q) attaching said next tentative labels onto said identified elements of said next candidate subset; thereafter

r) employing said attached next tentative labels via one of the class of inductive operations to formulate next local regularities; thereafter

s) tentatively identifying elements having specific combinations of said refined global regularities and said next local regularities to obtain attached next second tentative labels; thereafter

t) testing if estimated error rate is within a preselected tolerance or if a steady state in said next second tentative labels is evident; and if true, then rating confidence of said attached next second tentative labels and converting selected ones of said attached next second tentative labels to confidence labels upon achieving a preselected confidence level and then outputting data with said confidence labels;

u) otherwise employing said next second tentative labels via said operation on said candidate subset to formulate next second local regularities; and

v) repeating from step s).

14. The method according to claim 13 further including the steps of:

applying said data with confidence levels to further subsets of said dataset to investigate further subsets for local regularities.

15. In a data processing system, a method for learning and combining regularities of a first level and regularities of at least a second level and a third level for information extraction and classification, said method comprising the steps of:

a) determining a hierarchy of levels from most global level to most specific level;

b) beginning at the most global level, training a classifier at the selected level by initially providing descriptions of regularities at said selected level to a working database, said selected level regularities being patterns which may be found over a selected portion of a selected dataset; thereafter

c) identifying a candidate subset of the selected dataset in which next more specific regularities may be found; thereafter

d) tentatively identifying elements having said selected regularities in said candidate subset to obtain first tentative labels, said first tentative labels being useful for tagging like information; thereafter

e) attaching said first tentative labels onto said identified elements of said candidate subset; thereafter

f) employing said attached first tentative labels via one of a class of inductive operations to formulate first local regularities; thereafter

g) tentatively identifying elements having specific combinations of said global regularities and said local regularities to obtain attached second tentative labels; thereafter

h) testing if estimated error rate is within a preselected tolerance or if a steady state in said attached second tentative labels is evident; and if true, then rating confidence of said attached second tentative labels and converting selected ones of said attached second tentative labels to confidence labels upon achieving a preselected confidence level and then outputting data with said confidence labels;

i) otherwise employing said second tentative labels via said operation on said candidate subset to formulate second more specific regularities,

j) repeating from step g); and

k) repeating from Step b) for each successive more selective level of regularity.